Journal of Agricultural Machinery
Vol. 9, No. 2, Fall - Winter 2019, p. 469-484

نشریه ماشین‌های کشاورزی
جلد ۹، شماره ۲، نیمسال دوم ۱۳۹۸، ص ۴۸۴-۴۶۹

مدل‌سازی متغیرهای موثر بر عملکرد نیشکر با استفاده از الگوریتم‌های درخت تصمیم‌گیری C5.0 و QUEST

حسن ذکی دیزجی[1*]- هوشنگ بهرامی[2]- نسیم منجزی[3]- محمد جواد شیخ داودی[4]

تاریخ دریافت: ۱۳۹۶/۱۰/۰۵
تاریخ پذیرش: ۱۳۹۷/۰۱/۱۸

چکیده

در این پژوهش یکی از اهداف اصلی شرکت‌های کشت و صنعت نیشکر خوزستان که افزایش میزان عملکرد مزارع نیشکر با استفاده از رهیافت داده‌کاوی می‌باشد، مورد بررسی قرار گرفته است. تصمیم‌گیرندگان در این واحدهای تولیدی کشاورزی با حجم بسیار زیادی از داده‌های جمع‌آوری شده با خصوصیات بسیار متنوع و با روابط پیچیده در بین آن‌ها مواجه هستند که آنالیز و مدیریت آن‌ها به‌وسیله‌ی تجزیه و تحلیل‌های تجربی و آماری، امری دشوار و در بسیاری از حوضه‌ها عملاً ناممکن می‌باشد. داده‌کاوی یک فناوری توانمند در مدیریت و سازماندهی اطلاعات با حجم بالا می‌باشد. در این تحقیق با استفاده از تکنیک‌های داده‌کاوی درخت تصمیم (مدل‌های QUEST و C5.0)، به تخمین عملکرد محصول نیشکر پرداخته شده است. در این راستا مجموعه داده‌های در دسترس همچون داده‌های آبیاری و زهکشی، خاک و گیاه استفاده گردید تا اثر ترکیب‌های متفاوت این عوامل بر عملکرد تولید تعیین گردد. این پژوهش از نوع تحلیلی بوده و پایگاه داده آن شامل رکوردهای ۱۲۰۱ مزرعه می‌باشد. داده‌های مورد نیاز این تحقیق، طی سال‌های زراعی ۱۳۹۳ تا ۱۳۹۶ از کشت و صنعت امیرکبیر به‌دست آمده است. تجزیه و تحلیل به کمک نرم‌افزار IBM modeler 14.2 انجام شده است. نتایج نشان داد، شاخص‌های اجرایی و مدیریتی بر تغییر سطح عملکرد مزارع نیشکر تاثیرگذار می‌باشد. چگونگی تاثیرپذیری سطح عملکرد وابسته به ترکیب‌های خاصی از شاخص‌های اجرایی و مدیریتی می‌باشد که در قالب الگوهای حاصل از مدل‌های درخت تصمیم QUEST و C5.0 استخراج شده است. همچنین واریته محصول در هر دو مدل درخت تصمیم به‌عنوان مهم‌ترین متغیر مستقل در مدل‌سازی ظاهر شده است. بنابراین نتایج به‌دست آمده می‌تواند در برنامه‌ریزی و آماده‌سازی شرایط مطلوب برای رسیدن به اهداف تعیین شده میزان تولید کمک نماید.

واژه‌های کلیدی: ارزیابی، داده‌کاوی، کشاورزی، کشت و صنعت امیرکبیر

مقدمه

نیشکر، محصولی با ارزش و از منابع تأمین‌کننده ساکارز می‌باشد. گرچه از مهمترین اهداف کشاورزی نیشکر در بسیاری از کشورها تولید ساکارز، به‌عنوان یک ماده غذایی پرارزش است، اما در بعضی از کشورها که با مشکل تأمین منابع انرژی مواجه‌اند، کشت نیشکر به‌عنوان منبع مهم تولید انرژی نیز مطرح می‌باشد. برداشت حدودا ۱۲۵ تن در هکتار ساقه در سال از بعضی مزارع نیشکر خوزستان نشانگر استعداد اقلیمی نواحی گرم این استان برای کشت نیشکر می‌باشد. نیشکر در استان خوزستان به‌صورت صنعتی در قالب طرح توسعه نیشکر و صنایع جانبی استان خوزستان (کشت و صنعت‌های امام خمینی (ره)، امیرکبیر، میرزا کوچک خان، دعبل خزاعی، سلمان فارسی، فارابی و دهخدا) و کشت و صنعت‌های هفت تپه، کارون و میان آب در ابعاد جغرافیایی وسیعی (۱۱۰ هزار هکتار) کشت شده و مورد بهره‌برداری قرار می‌گیرد (Monjezi *et al.*, 2017). اما متاسفانه عملکرد نیشکر در طی سالیان گذشته رو به افول نهاده است. برخی از دلایل این افت عملکرد عبارتند از: فشردگی خاک مزارع در اثر رفت و آمد زیاد ماشین‌های سنگین زراعی (ماشین برداشت و سبد حمل نیشکر)، شوری آب رودخانه کارون، گرد و غبار و آلودگی هوای استان خوزستان. لذا به‌منظور بررسی علل کاهش عملکرد نیشکر و ارائه راهکارهایی به منظور افزایش تولید، نیاز به مطالعات گسترده‌ای در این زمینه می‌باشد. فرآیند داده‌کاوی در حال حاضر از راهکارهای موثر در تولید ارزش افزوده، استخراج دانش و مدیریت دانش می‌باشد. داده‌کاوی عبارتست از فرآیند یافتن دانش از مقادیر عظیم داده‌های ذخیره شده در پایگاه داده، انباره داده و یا دیگر مخازن اطلاعات. از پرکاربردترین روش‌های داده‌کاوی، درخت تصمیم است. درخت‌های

۱- استادیار گروه مهندسی بیوسیستم، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، اهواز، ایران
۲- دانشیار گروه مهندسی بیوسیستم، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، اهواز، ایران
۳- استادیار گروه مهندسی بیوسیستم، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، اهواز، ایران
۴- استاد گروه مهندسی بیوسیستم، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، اهواز، ایران
(*- نویسنده مسئول: Email: hzakid@scu.ac.ir)
DOI: 10.22067/jam.v9i2.69712

تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است. مدل‌های مختلف درخت تصمیم شامل Quest, CART,CHAID و C5.0 می‌باشد. در زمینه به‌کارگیری تکنیک‌های داده‌کاوی در کشاورزی تاکنون تحقیقات بسیاری انجام شده است که در ادامه به نمونه‌هایی از آن اشاره می‌شود.

Ramesh and Vardhan (2013) عملکرد محصولات کشاورزی را با استفاده از تکنیک‌های مختلف داده‌کاوی پیش‌بینی کردند. نتایج تحقیقات آن‌ها نشان داد که الگوریتم‌های k-means، k-Nearest Neighbor، support vector machines و شبکه‌های عصبی مصنوعی برای پیش‌بینی عملکرد محصولات کشاورزی از دقت بالا و توانایی زیادی برخوردار می‌باشد. Jeysenthil *et al.* (2014) با استفاده از تکنیک خوشه‌بندی داده‌کاوی (k-means) یک سیستم پشتیبانی برای پایگاه داده‌های خاک مزارع نیشکر طراحی کردند. Noorzadeh *et al.* (2011) کارآیی روش خوشه‌بندی را برای خوشه‌بندی غلظت مس با استفاده از ۲۱۳ نمونه خاک در اراضی کشاورزی استان همدان بررسی کردند. Goktepe *et al.* (2005) با استفاده از ۱۲۰ نمونه خاک و روش K-means و خوشه‌بندی مرحله‌ی، خاک‌های منطقه آنتالیا را طبقه‌بندی کردند. نتایج این تحقیق نشان داد، روش خوشه‌بندی نتیجه مطلوب‌تری داشته است. مطالعات دیگری نیز توانمندی و لزوم به‌کارگیری تکنیک‌های داده‌کاوی در کشاورزی را تصدیق می‌نمایند (Kalpana *et al.*, 2014 a, b; Geetha, 2015; Sharma and Mehta, 2012; Yethiraj, 2012; Rajesh, 2011; Raorane and Kulkarni, 2013).

هدف از این تحقیق، بررسی تاثیر و رتبه‌بندی عوامل اجرایی و مدیریتی علاوه بر عوامل طبیعی بر افزایش عملکرد تولید نیشکر در مزارع تحت پوشش شرکت کشت و صنعت امیرکبیر خوزستان با استفاده از مدل‌های C5.0 و QUEST و ارائه راهکارهایی به‌منظور افزایش عملکرد می‌باشد.

مواد و روش‌ها

در این تحقیق از داده‌های سال‌های زراعی ۹۶-۱۳۹۳ مزارع کشت و صنعت امیرکبیر، یکی از واحدهای هفت‌گانه شرکت توسعه نیشکر و صنایع جانبی استفاده شد. این کشت و صنعت در ۴۵ کیلومتری جنوب اهواز و در غرب رودخانه کارون و شرق جاده اهواز به خرمشهر و در طول جغرافیایی ۴۸ درجه و ۱۲ دقیقه تا ۴۸ درجه و ۳۰ دقیقه و عرض جغرافیایی ۳۱ درجه و ۱۵ دقیقه تا ۳۱ درجه و ۴۰ دقیقه قرار گرفته است. این منطقه دارای میانگین بارندگی سالیانه ۱۴۷/۱ میلی‌متر، میانگین دمای روزانه هوا ۲۵ درجه سلسیوس، میانگین دمای خاک ۲۱/۲ درجه سلسیوس و دارای میانگین ارتفاع ۷ متر از سطح دریا می‌باشد (Monjezi and Zakidizaji, 2017). مساحت این واحد ۱۴۰۰۰ هکتار و پتانسیل سطح زیر کشت آن ۱۲۰۰۰ هکتار می‌باشد و مابقی کانال، جاده، ساختمان و کارخانه می‌باشد. این واحد دارای ۴۸۰ مزرعه ۲۵/۵ هکتاری است.

مراحل اجرای تحقیق

مرحله اول: شناخت وضع موجود و شناسایی داده‌ها

عمده وظایف این مرحله شامل: شناخت وضعیت موجود، مطالعه تجربیات و مقالات مرتبط، بررسی فرآیند ثبت داده‌ها و منابع داده و تعیین دامنه طرح بود. هدف اصلی این مرحله ایجاد شناخت صحیح و همه‌جانبه دغدغه‌های موجود، چالش‌ها، اهداف و محدودیت‌های موجود در اجرای این تحقیق بود. در پایان این مرحله درک کامل و فهم مشترکی از صورت مسئله و راه‌حل‌های آن بین محققان و کارشناسان کشاورزی کشت و صنعت ایجاد گردید. فعالیت‌های انجام شده در این مرحله نیز شامل: انجام مصاحبه با مدیران و کارشناسان معاونت‌های کشاورزی و صنعت، بازدید از واحدهای کشاورزی مرتبط با تحقیق، مطالعه مقالات و تجربیات موفق کاربرد داده‌کاوی در حوزه کشاورزی، ارزیابی منابع داده موجود در بخش‌های مختلف همچون میزان تولید نیشکر، آبیاری و زهکشی، هواشناسی، خاک، گیاه و کلیه موارد مرتبط با عملکرد مزارع و تعیین دامنه و حدود تحقیق بر اساس محدودیت‌های موجود در منابع داده بود. همچنین با ارزیابی منابع داده‌ای موجود در کشت و صنعت، محدودیت و دامنه اجرایی پژوهش تعیین شد. تایید این مرحله منجر به ایجاد نقشه راه مشترک محققین و کارشناسان شرکت کشت و صنعت نیشکر گردید.

مرحله دوم: آماده‌سازی، مدل‌سازی و ارزیابی

عمده وظایف این مرحله عبارتست از: یکپارچه‌سازی و آماده‌سازی داده‌ها، مدل‌سازی و ارزیابی مدل‌ها و استخراج الگوها و تفسیر نتایج به‌دست آمده برای حل مسئله. طی این مرحله، مطالعات توصیفی، آماده‌سازی داده‌ها و مدل‌سازی به منظور تعیین مدل پیش‌بینانه عملکرد و کشف قوانین مفید برای کنترل و برنامه‌ریزی تولید با هدف افزایش عملکرد انجام پذیرفت. همچنین فعالیت‌های انجام شده در این مرحله نیز شامل: تعیین نوع هدف مناسب بر اساس محدودیت‌های موجود در منابع داده، به‌طور مثال انتخاب شاخص تن در هکتار به‌جای نرخ رشد هفتگی یا اندازه ارتفاع نیشکر، تعیین مقیاس زمانی مناسب برای ایجاد رکوردهای اطلاعاتی بر اساس محدودیت‌های موجود در منابع داده، به‌طور مثال انتخاب واحد روز، هفته، ماه و سال، یکپارچه‌سازی منابع داده از قبیل داده‌های آبیاری و زهکشی، خاک، گیاه و داده‌های عملکرد، پاک‌سازی، آماده‌سازی و نرمال‌سازی داده‌ها جهت ساخت مدل، تهیه طرح آزمون مناسب به‌منظور ارزیابی کیفیت مدل‌های به‌دست آمده، استخراج الگوها از مدل‌ها و بررسی کیفی قوانین و الگوهای به‌دست آمده از نگاه

داده‌کاوی، ارزیابی قوانین و الگوهای به‌دست آمده از منظر علوم کشاورزی و متخصصین نیشکر و قابلیت تفسیرپذیری آن‌ها و ارائه راهکار به‌منظور افزایش عملکرد بود.

بررسی وضعیت عملکرد مزارع نیشکر

آمار توصیفی عملکرد مزارع نیشکر در سال‌های زراعی ۱۳۹۳ تا ۱۳۹۶ در جدول ۱ آمده است.

جدول ۱- آمار توصیفی عملکرد مزارع نیشکر در سال‌های زراعی ۱۳۹۳ تا ۱۳۹۶

Table 1- Descriptive statistics on the performance of sugar cane fields in 2014-2017

میزان عملکرد و بازدهی در سال The yield and efficiency in year	کیفی Qualitative	کمی Quantitative (ton ha^{-1})	مساحت Area (ha)	فراوانی نسبی (درصد) Relative frequency (percent)
1393	خیلی کم very little	45.2-27.5	544.8	4.09
	کم Low	62.8-45.3	5705.4	42.84
	متوسط Medium	80.5-62.9	5911.3	44.38
	زیاد High	98.1-80.6	1543.8	11.59
	خیلی زیاد too much	115.8-98.2	158.7	1.19
1394	خیلی کم very little	38.5-19.2	992.54	7.16
	کم Low	57.7-38.6	5498.67	39.66
	متوسط Medium	77.1-57.8	6439.62	46.45
	زیاد High	96.3-77.2	876.20	6.32
	خیلی زیاد too much	115.5-96.4	57.04	0.41
1395	خیلی کم very little	53.1-32.9	533.76	3.85
	کم Low	73.1-53.2	8211.61	59.23
	متوسط Medium	93.2-73.2	4435.54	31.99
	زیاد High	113.4-93.3	630.28	4.55
	خیلی زیاد too much	133.5-113.5	52.88	0.38
1396	خیلی کم very little	49.8-27.7	241.36	1.74
	کم Low	71.8-49.9	5303.25	38.25
	متوسط Medium	93.8-71.9	6879.89	49.62
	زیاد High	116.1-93.9	1365.84	9.85
	خیلی زیاد too much	138.1-116.2	73.72	0.53

توصیف داده‌ها

داده‌های استفاده شده در این تحقیق شامل ۱۵ متغیر می‌باشند که از ۱۲۰۱ مزرعه نیشکر در طی سال‌های ۹۶-۱۳۹۳ به‌دست آمده‌اند.

متغیرهای مورد استفاده به دو دسته متغیرهای پیش‌گویی‌کننده و متغیر هدف تقسیم شدند. متغیر عملکرد مزارع به‌عنوان متغیر هدف (متغیر وابسته) و سایر متغیرها به‌عنوان متغیر پیش‌گویی‌کننده (متغیر

مستقل) در نظر گرفته شدند. در هر دو مدل QUEST و C5.0 داده‌های ورودی شامل: واریته محصول، ماه برداشت، کود شیمیایی (نیتروژن)، کود شیمیایی (فسفر)، سن گیاه (کشت اول یا راتون)، تعداد دفعات آبیاری مزرعه در طی فصل زراعی، نسبت سطح سمپاشی مزرعه (نسبت سطح سمپاشی شده به مساحت کل مزرعه)، بافت خاک، هدایت الکتریکی خاک (EC)، مقدار مصرف آب در هکتار، زهکشی، مدیریت مزرعه (مدیریت تولید)، طول فصل زراعی، مساحت مزرعه و عملکرد نیشکر می‌باشد. اطلاعات و توصیف داده‌های ورودی در دو مدل QUEST و C5.0 در جداول ۲ و ۳ آمده است.

جدول ۲- توصیف متغیرهای پیوسته ورودی مدل‌ها

Table 2- Description of continuous sugarcane variables used for this study

نام متغیر Variable name	واحد Unit	نوع متغیر Variable's Type	نقش متغیر Usage (role)	توصیف متغیر Description کمترین مقدار Minimum amount	بیشترین مقدار Maximum amount	میانگین Average	انحراف معیار Standard deviation	رکوردهای در دسترس Number of valid records
زهکشی Drain	m^3ha^{-1}	پیوسته Continuous	ورودی Input	8211.91	29554.50	16893.62	3697.66	1201
مساحت Area	ha	پیوسته Continuous	ورودی Input	5	35	23.36	5.18	1201
کود شیمیایی (نیتروژن) Chemical fertilizer (Nitrogen)	kg ha^{-1}	پیوسته Continuous	ورودی Input	160	553	340.76	46.88	1201
کود شیمیایی (فسفر) Chemical fertilizer (Phosphate)	kg ha^{-1}	پیوسته Continuous	ورودی Input	0	250	64.55	99.86	1201
تعداد دفعات آبیاری Times irrigation	-	پیوسته Continuous	ورودی Input	12	34	21.29	3.95	1201
هدایت الکتریکی خاک Soil electrical conductivity (EC)	ds m^{-1}	پیوسته Continuous	ورودی Input	2	16	5.73	2.13	1201
مقدار آب مصرفی Water consumption per hectare	m^3ha^{-1}	پیوسته Continuous	ورودی Input	1005	1900	1395.55	133.58	1201
طول فصل زراعی Crop duration	day	پیوسته Continuous	ورودی Input	389	567	443.37	32.32	1201

درخت تصمیم

درخت‌های تصمیم معمولا بوسیله نمودارهای ساقه و برگ نمایش داده می‌شوند. هر برگ در درخت تصمیم مشخص کننده یک طبقه می‌باشد؛ درحالی‌که، یک ساقه بیانگر شروط یک ویژگی است. درخت‌های تصمیم بخاطر کاربردشان در طبقه‌بندی‌های گسسته معروفند. این طبقه‌بندی گسسته به بیان منطقی به ارزش یک ویژگی وابسته است. علاوه بر این، از درخت‌های تصمیم می‌توان برای طبقه‌بندی تبیینی استفاده کرد؛ و همچنین الگوریتم‌های یادگیری آن‌ها سریع می‌باشند. الگوریتم درخت تصمیم قادر است علاوه بر متغیرهای کمی، متغیرهای کیفی را نیز پیش‌بینی کند. این روش اولین بار توسط Breman *et al.* (1984) ارائه شد. الگوریتم درخت تصمیم به گونه‌ای عمل می‌کند که سعی دارد گوناگونی و یا تنوع (از نظر ویژگی هدف) را در گره‌ها به حداقل ممکن برساند. این عدم یکنواختی در گره‌ها با استفاده از معیارهای عدم خلوص[۱] قابل اندازه‌گیری است که مهمترین و پرکاربردترین آن شاخص جینی[۲] می‌باشد (Yoneyama *et al.*, 2002).

اغلب تفاوت انواع درخت‌های تصمیم در همین معیار اندازه‌گیری عدم خلوص، شیوه شاخه‌بندی[۳] و هرس کردن گره‌های درخت می‌باشد. در این پژوهش از دو نوع الگوریتم درخت تصمیم C5.0 و QUEST استفاده شد.

1- Impurity measure
2- Gini
3- Splitting

جدول ۳- توصیف متغیرهای گسسته ورودی مدل‌ها

Table 3- Description of categorical sugarcane variables used for this study

نام متغیر **Variable name**	نوع متغیر **Variable's Type**	نقش متغیر **Usage (role)**	توصیف متغیر **Description**	تعداد مزارع موجود در این گروه **n[a]**	مساحت **Area (%)**
مدیریت مزرعه Farm management	گسسته Categorical	ورودی Input	تولید اول (۶۰۰۰ هکتار) First Production Manager (6000 ha)	598	49.79
			تولید دوم (۶۰۰۰ هکتار) Second Production Manager (6000 ha)	603	50.21
واریته محصول Crop cultivar	گسسته Categorical	ورودی Input	SP70-1143	234	19.48
			IRC99-02	22	1.83
			CP69-1062	619	51.54
			CP57-614	153	12.74
			CP48-103	173	14.40
بافت خاک Soil texture	گسسته Categorical	ورودی Input	Sandy loam شنی لوم	76	6.33
			Sandy clay loam شنی رسی لوم	36	3.00
			Silt loam سیلتی لوم	24	2.00
			Silt clay سیلتی رسی	74	6.16
			Silt clay loam سیلتی کلی لوم	255	21.23
			Loam لوم	414	34.47
			Clay loam کلی لوم	322	26.81
سن گیاه Age	گسسته Categorical	ورودی Input	کشت اول Plant	359	29.89
			راتون اول First ratoon	358	29.81
			راتون دوم Second ratoon	293	24.40
			راتون سوم Third ratoon	191	15.90
نسبت سطح سمپاشی ratio of surface spraying [b]	گسسته Categorical	ورودی Input	1	271	22.56
			2	371	30.89
			3	296	24.65
			4	263	21.90
عملکرد Yield (ton ha^{-1})	گسسته Categorical	خروجی (هدف) Target	مطلوب (۱۰۰≤) High (≥100)	90	7.49
			متوسط (۱۰۰>x≥۶۵) Medium (65≤x<100)	541	45.05
			نامطلوب (۶۵>) Low (<65)	570	47.46
ماه برداشت Month of harvest	گسسته Categorical	ورودی Input	(مهر، اکتبر) ۱۰ 10 (October)	81	6.74
			(آبان، نوامبر) ۱۱ 11 (November)	190	15.82
			(آذر، دسامبر) ۱۲ 12 (December)	328	27.31
			(دی، ژانویه) ۱ 1 (January)	128	10.66
			(بهمن، فوریه) ۲ 2 (February)	262	21.82
			(اسفند، مارس) ۳ 3 (March)	172	14.32
			(فروردین، آوریل) ۴ 4 (April)	20	1.67
			(اردیبهشت، مه) ۵ 5 (May)	20	1.67

مدل C5.0

این الگوریتم بهبود یافته درخت‌های C4.5 و ID3 می‌باشد (Quinlan, 1993). تقسیم‌بندی هر گره بر اساس بهره اطلاعاتی محاسبه می‌گردد. این شاخص برای انتخاب متغیر شکننده در فرآیند رشد درخت مورد استفاده قرار می‌گیرد. همگنی نمونه‌ها در یک گره با شاخص آنتروپی معین می‌شود. در جهت محاسبه بهره اطلاعاتی ابتدا باید آنتروپی محاسبه گردد. اگر متغیر هدف دارای C مقدار متفاوت باشد، آنگاه آنتروپی S وابسته به C کلاس از رابطه (۱) حاصل می‌شود (Kotsiantis, 2007).

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \qquad (۱)$$

که p_i نسبتی از S می‌باشد که به کلاس i تعلق دارد. بهره اطلاعاتی میزان کاهش مورد انتظار در آنتروپی را نشان می‌دهد. آنتروپی خلوص داده‌ها در یک گزینه را ارائه می‌دهد و بهره اطلاعاتی تاثیر یک متغیر در کلاس‌بندی را معین می‌سازد. بهره اطلاعاتی (S,A) مربوط به متغیر A وابسته به داده‌های S با رابطه (۲) محاسبه می‌گردد (Kotsiantis, 2007):

$$Gain(S,A) = Entropy(S) - \sum_{V\epsilon Value(A)} \frac{|S_V|}{|S|} Entropy(S_v) \qquad (۲)$$

که در آن Value (A) تمامی مقادیر ممکن متغیر A می‌باشد و S_V زیر مجموعه‌ای از S می‌باشد که برای متغیر A دارای مقدار V است. قسمت اول رابطه بالا مربوط به آنتروپی S در حالت اولیه و قسمت دوم آنتروپی مورد انتظار پس از تقسیم بر اساس متغیر A است. در هر شاخه رشد یافته درخت هر متغیر تنها یک بار حضور می‌یابد. رشد درخت تا مرحله‌ای که همه متغیرها در یک شاخه حضور یابند یا تمامی نمونه‌ها در یک گره متغیر به یک دسته باشند، ادامه می‌یابد.

مدل QUEST

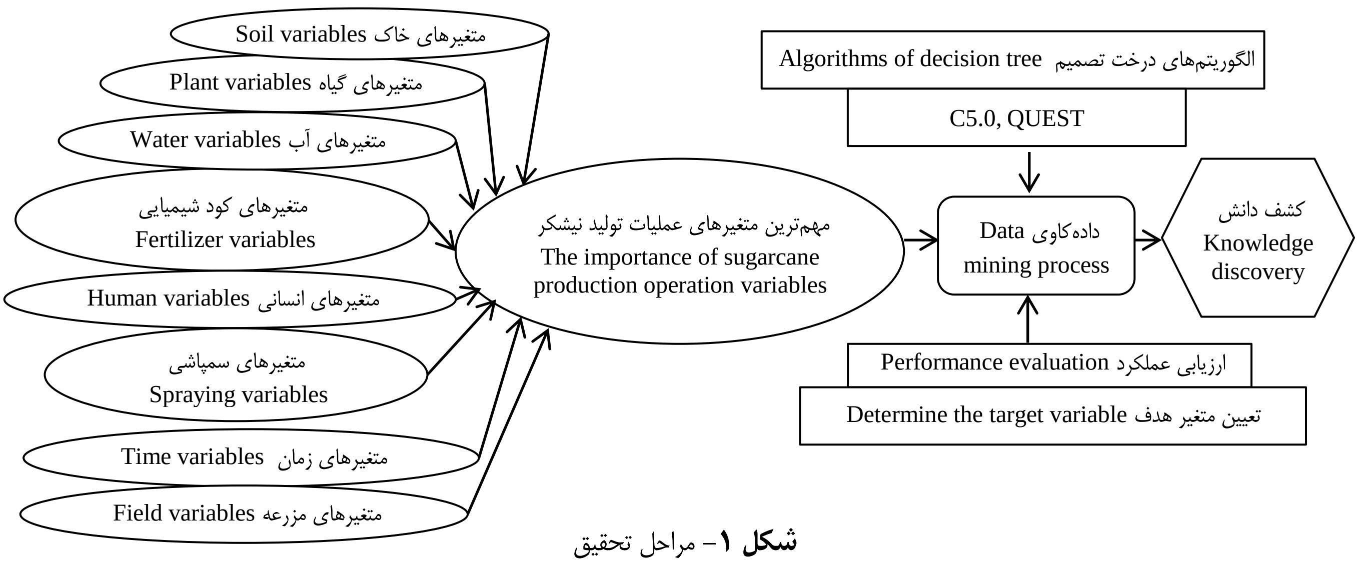
این روش، دسته‌بندی دوگانه را برای ساخت یک درخت ارائه می‌دهد. این تکنیک با هدف کاهش زمان مورد نیاز برای ساخت درخت و کاهش تمایلی که با وجود متغیرهای توصیفی پیوسته، در جواب‌های درخت حاصل می‌شود، به‌وجود آمده است. QUEST از قانون‌هایی متناوب بر مبنای آزمون معناداری برای ارزیابی متغیر توصیفی تقسیم‌کننده‌گر استفاده می‌کند. همگنی داده‌ها در هر گره بر مبنای نسبت واریانس داخلی دسته و واریانس میان دسته‌ها، با آماره F و از رابطه (۳) محاسبه می‌گردد:

$$F_X = \frac{\sum_{C=1}^{C=1} N_c(t)\frac{(\bar{x}_C(t)-\bar{x}(t)))^2}{C-1}}{\sum_{C=1}^{C=1}\frac{(x_i-\bar{x}_C(t)))}{(n(t)-C)}} \sim F(c-a; n(t)-c) \qquad (۳)$$

که در آن $\bar{X}_C\ (t)$ میانگین متغیر توصیفی X در گروه c برای گره t و x(t) میانگین متغیر توصیفی x در گروه t برای همه گره‌ها می‌باشد (Razi Ardakani and samimi, 2011).

تجزیه و تحلیل داده‌ها

در این تحقیق از نرم‌افزار داده‌کاوی IBM SPSS Modeler 14.2 برای مدل‌سازی درخت‌های تصمیم و اعتبارسنجی نتایج بهره گرفته شد. داده‌ها به دو دسته داده‌های آموزش و داده‌های آزمایش تقسیم شدند که سهم داده‌های آموزشی ۷۰ درصد داده‌ها و سهم داده‌های آزمایش ۳۰ درصد داده‌ها بود. مراحل تحقیق در شکل ۱ آورده شده است.



شکل ۱- مراحل تحقیق

Fig.1. Proposed Framework

نتایج و بحث

مدل‌سازی

در اینجا به منظور بررسی روابط بین شاخص‌های اجرایی و سطح عملکرد از مدل‌های مبتنی بر تولید قانون استفاده شده است. بنابراین انواع مدل‌های درخت تصمیم QUEST و C5.0 جهت افراز مجموعه مزارع مورد بررسی به زیر مجموعه‌های کوچکتر و همگن که دارای سطوح عملکردی مشابه و نزدیک به هم می‌باشند، اجرا شده است.

تحلیل مدل درخت تصمیم QUEST

در این الگوریتم ۷۰/۰۲ درصد از داده‌ها برای آموزش مدل و ۲۹/۹۸ درصد داده‌ها برای آزمایش مدل استفاده شدند. سطح معنی‌داری برای فاکتور تقسیم‌کننده برابر ۰/۰۵ بود. همچنین درخت تصمیم به‌دست آمده از این مدل دارای ۵ سطح می‌باشد. شکل ۲ درخت تصمیم مدل QUEST را برای عملکرد محصول نیشکر نمایش می‌دهد. در ساختار درخت تصمیم هر گره داخلی برای نمایش آزمایش بر روی یک متغیر و هر شاخه از درخت نشان‌دهنده نتایج خروجی آزمایش می‌باشند. همچنین هر گره برگ از درخت یک سطح از طبقه‌بندی و دسته‌بندی متغیر را نشان می‌دهد (تصمیم نهایی بعد از محاسبه تمامی متغیرها گرفته می‌شود). در درخت تصمیم هر مسیر از ریشه تا برگ انتهایی نشان‌دهنده قوانین کلاس‌بندی می‌باشد. درخت تصمیم الگوریتم QUEST در شکل ۲ نشان‌دهنده یک درخت تصمیم با ۸ گره می‌باشد که ۵ گره از آن‌ها، گره‌های نهایی هستند. شاخص کای اسکوئر[1] به‌عنوان یک معیار تقسیم‌کننده درخت انتخاب شده است. عملکرد محصول نیشکر به‌عنوان ریشه درخت توسط واریته IRC99-02 و واریته‌های CP48-103, CP57-614, CP69-1062, SP70-1143 به گره و شاخه‌های جدید تقسیم شده است و ادامه درخت به همین صورت گسترش یافته است. مدل درخت تصمیم QUEST برای عملکرد نیشکر بیانگر تاثیر بالای واریته محصول است. نتایج حاصل از درخت تصمیم قابل گسترش و ارائه در قالب قوانین اگر– آنگاه[2] می‌باشند (شکل ۳).

تحلیل مدل درخت تصمیم C5.0

در این مدل ۷۰/۰۲ درصد از داده‌های موجود در تحقیق برای آموزش مدل و ۲۹/۹۸ درصد از این داده‌ها برای آزمایش مدل استفاده شده است. داده‌های آموزشی در مدل C5.0 سبب استخراج قوانین کلاس‌بندی شد. سپس زمانی که مدل C5.0 ساخته شد، این قوانین برای کلاس‌بندی داده‌های آزمایش به‌کار گرفته شدند. همچنین مجموعه‌ای از قوانین اگر–آنگاه توسط مدل جمع‌آوری شده است. هرکدام از این مجموعه قوانین استخراجی، رابطه بین متغیرها و کلاس‌بندی انجام شده را نمایش می‌دهند. در اینجا قسمت "اگر" قانون به‌عنوان یک پیش‌شرط شناخته می‌شود و قسمت "آنگاه" قانون به‌عنوان نتیجه و پیامد قانون به حساب می‌آید. پشتیبانی[3] یک قانون، تعداد گره‌های قانون قبلی است. دقت کلی مدل شامل نسبت رکوردهای درست پیش‌بینی شده به کل رکوردها است. دقت[4] هر قانون، نسبت رکوردهای درست پیش‌بینی شده یک قانون به رکوردهای پیش‌بینی شده در آن قانون است. ساپورت[5] هر قانون نشان‌دهنده تعداد رکوردهای آن قانون است. اطمینان[6] قانون نیز نسبتی از رکوردهای هر قانون است که به درستی پیش‌بینی شده‌اند. ارتقاء[7] قانون نسبت اطمینان هر قانون به احتمال اولیه آن کلاس است. در این مدل، ۲۴ مجموعه قانون کلاس‌بندی برای پیش‌بینی عملکرد محصول نیشکر به‌دست آمده است. این قوانین کلاس‌بندی می‌توانند به سادگی برای مزارع جدید استفاده شوند. لیست قوانین استخراجی مدل C5.0 در جدول ۴ آمده است. مقدار پشتیبانی و دقت مدل درون پرانتز برای هر قانون مشخص گردیده است.

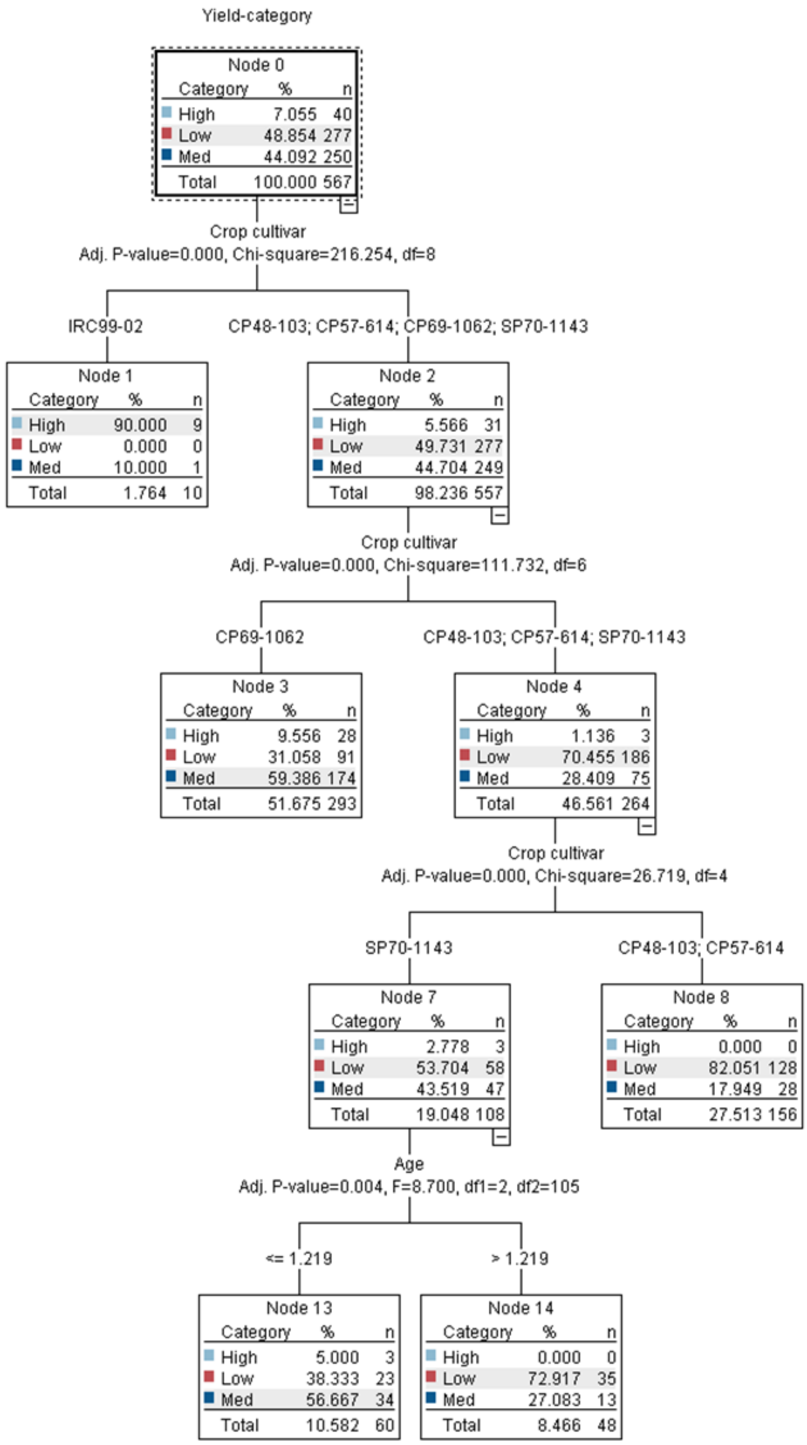
شاخص اهمیت متغیرها در مدل‌سازی

یکی دیگر از خروجی‌های مدل درخت تصمیم، رتبه‌بندی اهمیت متغیرها است. معمولاً در مدل‌سازی بر متغیرهایی که اهمیت بیشتری دارند تمرکز می‌شود و متغیرهای کم اهمیت‌تر حذف می‌شوند. در مدل QUEST متغیر واریته محصول و سن گیاه به‌ترتیب به‌عنوان متغیرهای با اهمیت در پیش‌بینی عملکرد محصول نیشکر به‌دست آمدند. کم‌اهمیت‌ترین متغیر برای پیش‌بینی عملکرد محصول نیشکر نیز متغیر نسبت سطح سمپاشی است. بر اساس الگوریتم C5.0 نیز، واریته محصول به‌عنوان با اهمیت‌ترین متغیر شناخته شده است و در جایگاه بعدی به‌ترتیب متغیرهای تعداد دفعات آبیاری و سن گیاه قرار دارند. سن گیاه فاکتور تاثیرگذاری بر عملکرد نیشکر به حساب می‌آید. نیشکر گیاهی چند ساله است و بهره‌برداری از آن به یک سال ختم نمی‌شود. از سوی دیگر عملکرد نیشکر در سال اول (مرحله کشت اول) معمولاً در حداکثر مقدار است و در سال‌های بهره‌برداری بعدی (راتون‌های ۱، ۲، ۳ و غیره) به‌تدریج از عملکرد آن کاسته می‌شود. در مجموع با توجه به نتایج حاصل از هر دو مدل به راحتی می‌توان به این نتیجه رسید که متغیر واریته محصول به‌عنوان مهم‌ترین و با اهمیت‌ترین متغیر در پیش‌بینی عملکرد محصول نیشکر نقش دارد. جدول ۵، شاخص اهمیت متغیرها در پیش‌بینی محصول نیشکر در هر دو مدل QUEST و C5.0 نشان داده شده است. همان‌طور که از نتایج بر می‌آید متغیر واریته محصول در هر دو مدل به‌عنوان

1- Chi-square index
2- IF-THEN rules

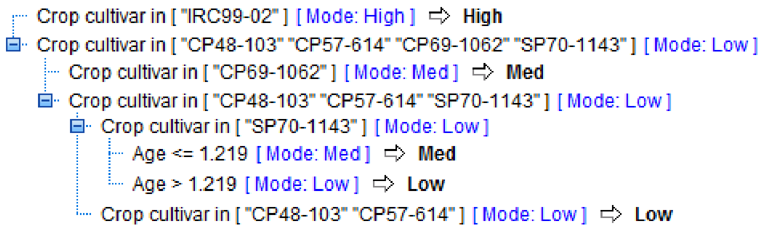
3- Coverage
4- Accuracy
5- Support
6- Confidence
7- Lift

مهم‌ترین متغیر مستقل در عملکرد محصول نیشکر نقش‌آفرینی می‌کند.



شکل ۲– مدل درخت تصمیم QUEST برای پیش‌بینی عملکرد محصول نیشکر

Fig. 2. QUEST model predicting yield of sugarcane



شکل ۳– قوانین استخراجی از الگوریتم درخت تصمیم QUEST

Fig. 3. Rules obtained from the application of QUEST decision tree algorithm

جدول ٤– مجموعه قوانین استخراجی درخت تصمیم مدل C5.0

Table 4- The result of C5.0 decision tree rule

Rule set generated by C5.0 algorithm (decision making support*)
Rules for High - contains 4 rule(s) Rule 1 for High (9; 0.909) if Chemical fertilizer (Phosphate) > 215 and Times irrigation <= 20 and Water consumption per hectare <= 1,444 then High Rule 2 for High (19; 0.81) if Crop cultivar = IRC99-02 then High Rule 3 for High (3; 0.8) if Crop cultivar = SP70-1143 and Age <= 1 and Chemical fertilizer (Phosphate) <= 215 and Chemical fertilizer (Nitrogen) > 368 and ratio of surface spraying = 2.000 then High Rule 4 for High (5; 0.714) if Crop cultivar = CP69-1062 and Soil electrical conductivity (EC) <= 4 and Chemical fertilizer (Nitrogen) > 425 and Times irrigation <= 19 then High
Rules for Low - contains 8 rule(s) Rule 1 for Low (51; 0.906) if Area > 19 and Crop cultivar = SP70-1143 and Age > 1 and Month of harvest in [1.000 8.000 9.000 10.000 11.000 12.000] then Low Rule 2 for Low (183; 0.886) if Area > 20 and Crop cultivar in ["CP48-103" "CP57-614"] then Low Rule 3 for Low (157; 0.855) if Crop cultivar in ["CP48-103" "CP57-614"] and Soil electrical conductivity (EC) > 4 then Low Rule 4 for Low (4; 0.833) if Area > 29 and Soil electrical conductivity (EC) > 5 and Chemical fertilizer (Phosphate) <= 215 then Low Rule 5 for Low (4; 0.833)

if Crop cultivar = SP70-1143
and Drain <= 10565.200
then Low
Rule 6 for Low (26; 0.786)
if Farm management = First Production Manager
and Chemical fertilizer (Nitrogen) <= 425
and Times irrigation <= 16
and ratio of surface spraying in [2.000 3.000 4.000]
then Low
Rule 7 for Low (81; 0.687)
if Area <= 24
and Soil electrical conductivity (EC) > 5
and Chemical fertilizer (Phosphate) <= 215
then Low
Rule 8 for Low (119; 0.645)
if Crop cultivar = SP70-1143
and Chemical fertilizer (Phosphate) <= 195
and Chemical fertilizer (Nitrogen) <= 368
then Low

Rules for Med - contains 12 rule(s)
Rule 1 for Med (10; 0.917)
if Age <= 1
and Soil texture = Silt loam
then Med
Rule 2 for Med (10; 0.917)
if Crop cultivar = SP70-1143
and Age <= 1
and Chemical fertilizer (Phosphate) <= 215
and Chemical fertilizer (Nitrogen) > 368
and ratio of surface spraying in [1.000 3.000 4.000]
then Med
Rule 3 for Med (8; 0.9)
if Crop cultivar = SP70-1143
and Chemical fertilizer (Phosphate) > 195
and Chemical fertilizer (Phosphate) <= 215
then Med
Rule 4 for Med (7; 0.889)
if Crop cultivar = CP69-1062
and Soil electrical conductivity (EC) > 4
and Chemical fertilizer (Nitrogen) > 425
then Med
Rule 5 for Med (7; 0.889)
if Crop cultivar = SP70-1143
and Age <= 1
and Soil texture = Loam
and Water consumption per hectare > 1481.100
then Med
Rule 6 for Med (7; 0.889)
if Crop cultivar = SP70-1143
and Age > 1
and Soil electrical conductivity (EC) <= 7
and Month of harvest in [2.000 7.000]
then Med
Rule 7 for Med (5; 0.857)
if Age <= 1
and Farm management = First Production Manager
and Soil texture = Sandy loam
and Chemical fertilizer (Phosphate) <= 215
and ratio of surface spraying in [1.000 2.000]

```
            then Med
    Rule 8 for  Med (156; 0.772)
            if Crop cultivar = CP69-1062
            and Soil electrical conductivity (EC) <= 5
            and Chemical fertilizer (Phosphate) <= 215
            and Chemical fertilizer (Nitrogen) <= 425
            and Times irrigation > 16
            then Med
    Rule 9 for  Med (54; 0.679)
            if Area <= 20
            and Soil electrical conductivity (EC) <= 4
            then Med
    Rule 10 for  Med (43; 0.667)
            if Area > 21
            and Crop cultivar in [ "CP69-1062" "SP70-1143" ]
            and Age <= 1
            and Soil texture = Silt clay loam
            and Chemical fertilizer (Phosphate) <= 195
            then Med
    Rule 11 for  Med (241; 0.658)
            if Area > 24
            and Area <= 29
            and Crop cultivar = CP69-1062
            and Chemical fertilizer (Phosphate) <= 215
            then Med
    Rule 12 for  Med (96; 0.612)
            if Crop cultivar in [ "CP69-1062" "SP70-1143" ]
            and Chemical fertilizer (Phosphate) > 215
            and Times irrigation > 20
            then Med
```

*Instance and confidence figures

جدول ۵– مقادیر ضرایب نرمال شده متغیرها

Table 5- The normalized importance of variables in classification

متغیر **Variable**	واریته محصول **Crop cultivar**	سن گیاه **Age**	تعداد دفعات آبیاری **Times irrigation**	زهکشی **Drain**	مساحت مزرعه **Area**	کود شیمیایی (نیتروژن) **Chemical fertilizer (Nitrogen)**	کود شیمیایی (فسفر) **Chemical fertilizer (Phosphate)**	هدایت الکتریکی خاک **Soil electrical conductivity (EC)**	مصرف آب در هکتار **Water consumption per hectare**	طول فصل زراعی **Crop duration**	ماه برداشت **Month of harvest**	مدیریت مزرعه **Farm management**	بافت خاک **Soil texture**	نسبت سطح سمپاشی **Ratio of surface spraying**
اهمیت (مدل C5.0) Importance (C5.0 algorithm)	0.10	0.077	0.081	0.071	0.074	0.076	0.074	0.070	0.072	0	0.075	0.074	0.077	0.073
اهمیت (مدل QUEST) Importance (QUEST algorithm)	0.762	0.057	0.018	0.018	0	0.018	0.018	0	0.018	0.018	0.018	0.018	0.018	0.018

ارزیابی مدل (دقت مدل)

نرخ پیش‌بینی اندازه‌گیری شده برای مدل C5.0 برابر ۷۶/۸۱ درصد پیش‌بینی‌های درست و ۲۳/۱۹ درصد پیش‌بینی‌های غلط برای داده‌های بخش آموزش و ۶۶/۹۴ درصد پیش‌بینی درست و ۳۳/۰۶ درصد پیش‌بینی غلط برای داده‌های آزمایش می‌باشد. همچنین در مدل QUEST نرخ پیش‌بینی درست برای داده‌های آموزش ۶۸/۲۵ درصد و نرخ پیش‌بینی غلط ۳۱/۷۵ درصد است و برای داده‌های تست نرخ پیش‌بینی درست ۷۰/۸۳ درصد و برای پیش‌بینی‌های غلط

۲۹/۱۷ درصد می‌باشد (جدول ۶). نتایج حاصل از مقایسه داده‌های آموزش بین دو مدل نشان می‌دهد که مدل C5.0 عملکرد بهتری در پیش‌بینی عملکرد محصول نیشکر از خود نشان می‌دهد. همچنین در داده‌های آموزش و مقایسه بین دو مدل به این نکته پی برده می‌شود که مدل QUEST دارای عملکرد بهتری در این زمینه می‌باشد. بنابراین به‌طور کلی، نتایج ارزیابی مدل‌ها اختلاف معناداری را بین عملکرد دو مدل نشان نمی‌دهد. تحقیقات دیگری که در زمینه کاربرد مدل های C5.0 و QUEST انجام شده است، سطح موفقیت مختلفی در این زمینه را گزارش داده‌اند. به‌عنوان نمونه در تحقیق Choi *et al.* (2014) نرخ پیش‌بینی اندازه‌گیری شده برای مدل C5.0، ۹۴/۶۴ درصد پیش‌بینی درست و ۵/۳۶ درصد پیش‌بینی غلط برای داده‌های آموزش و ۸۰/۷۷ درصد پیش‌بینی درست و ۱۹/۲۳ درصد پیش‌بینی غلط برای داده‌های آزمایش است. نرخ پیش‌بینی اندازه‌گیری شده برای مدلQUEST، ۵۰ درصد پیش‌بینی درست و ۵۰ درصد پیش‌بینی غلط برای داده‌های آموزش و ۴۱/۶۷ درصد پیش‌بینی درست و ۵۸/۳۳ درصد پیش‌بینی غلط برای داده‌های آزمایش است. در تحقیقی دیگر، نرخ پیش‌بینی اندازه‌گیری شده برای مدل C5.0، ۷۵/۵۲ درصد پیش‌بینی درست و ۲۴/۴۸ درصد پیش‌بینی غلط برای داده‌های آزمایش است (Umesh and Thilak, 2015). در تحقیق Ekasingh and Ngamsomsuke (2009)، نرخ پیش‌بینی اندازه‌گیری شده برای مدل C5.0، ۸۴ درصد پیش‌بینی درست و ۱۶ درصد پیش‌بینی غلط برای داده‌های آزمایش است. همچنین Baisen and Tillman (2007)، پتانسیل استفاده از مدل‌های درخت تصمیم را در مدل‌سازی کارآیی استفاده از کود نیتروژن در چراگاه‌های نیوزیلند بررسی کردند. ایشان تایید کردند که دقت مدل آن‌ها برای ۱۱ آزمون انجام شده برابر ۶۹ درصد بود.

جدول ٦- ارزیابی عملکرد الگوریتم‌های درخت تصمیم

Table 6- Performance evaluation of decision tree algorithms

مدل‌ها Models	داده‌های آموزش Training data		داده‌های تست Test data	
	درست٪ Correct%	غلط٪ Wrong%	درست٪ Correct%	غلط٪ Wrong%
C5.0	76.81	23.19	66.94	33.06
QUEST	68.25	31.75	70.83	29.17

نتیجه‌گیری

تحقیق حاضر، یکی از اولین مطالعات عارضه‌یابی عملکرد نیشکر در سطح کشور می‌باشد. مهم‌ترین نوآوری که این تحقیق می‌تواند به دنبال داشته باشد، برآورد دقیقی از میزان تاثیر عوامل مختلف تحت بررسی بر میزان عملکرد تن در هکتار مزارع نیشکر است. در این تحقیق، نیاز به به‌کارگیری روش‌های نوین در ساماندهی پایگاه‌های بزرگ اطلاعات عملکرد نیشکر و نیز سودمندی روش‌های داده‌کاوی به‌خصوص درخت تصمیم در تخمین عملکرد با حداکثر استفاده از دیگر پارامترها و اطلاعات برداشت شده، مورد بحث و بررسی قرار گرفت. از داده‌های شرکت کشت و صنعت نیشکر امیرکبیر استفاده شد. پایگاه داده تهیه شده به دو بخش آموزشی و تست، تقسیم گردید. در ادامه، مدل‌هایی بر اساس تکنیک درخت تصمیم داده‌کاوی برای تخمین عملکرد محصول نیشکر با استفاده از نرم‌افزار IBM modeler 14.2 و مجموعه داده‌های آموزشی، تولید گردید. در نهایت مدل تولید شده با استفاده از مجموعه داده‌های تست، مورد ارزیابی قرار گرفت. نتایج این پژوهش نشان داد که از تکنیک‌های داده‌کاوی از جمله مدل‌های درختی می‌توان در برآورد عملکرد محصول نیشکر استفاده نمود. بر اساس نتایج ارائه شده شاخص‌های اجرایی و مدیریتی بر تغییر سطح عملکرد تاثیرگذار می‌باشد. چگونگی تاثیرپذیری سطح عملکرد وابسته به ترکیب‌های خاصی از شاخص‌های اجرایی و مدیریتی می‌باشد که در قالب الگوهای حاصل از مدل درخت تصمیم QUEST و C5.0 استخراج شده است. نتایج حاصل از مقایسه داده‌های آموزش بین دو مدل نشان می‌دهد که مدل C5.0 عملکرد بهتری در پیش‌بینی عملکرد محصول نیشکر از خود نشان می‌دهد. همچنین در داده‌های آموزش و مقایسه بین دو مدل به این نکته پی برده می‌شود که مدل QUEST دارای عملکرد بهتری در این زمینه می‌باشد. بنابراین به‌طور کلی، نتایج ارزیابی مدل‌ها اختلاف معناداری را بین عملکرد دو مدل نشان نمی‌دهد. بنابراین با توجه به سطح زیر کشت بالای نیشکر در خوزستان می‌توان اطلاعات حاصل از این پردازش را در اختیار مدیران تولید این واحدهای کشت و صنعتی قرار داد و نتایج به‌دست آمده می‌تواند در برنامه‌ریزی و آماده‌سازی شرایط مناسب برای رسیدن به اهداف تعیین شده میزان تولید در شرکت‌های کشت و صنعت نیشکر کمک نماید.

پیشنهادها

با توجه به پتانسیل بالای ذخیره‌سازی و تحلیل پیشرفته داده‌ها در شرکت‌های کشت و صنعت نیشکر خوزستان موارد زیر پیشنهاد می‌گردد:

- جمع‌آوری سیستماتیک و مکانیزه داده‌ها با ایجاد زیرساخت نرم‌افزاری مناسب

- ایجاد انبار داده با هدف استفاده در ابزارهای گزارش‌گیری و تحلیلی
- استقرار سیستم هوشمند کسب و کار (BI) به همراه تعریف و اجرای پروژه‌های داده‌کاوی

همچنین پیشنهاد می‌شود تا در مطالعات آینده از سایر تکنیک‌های موجود در داده‌کاوی استفاده شود و با نتایج این پژوهش مقایسه شود. همچنین تبین قوانینی برای درج دقیق متغیرهای ورودی و خروجی در پایگاه داده توسط شرکت کشت و صنعت می‌تواند کمک شایانی در امر داده‌کاوی باشد تا بتوان نتایج واقعی‌تری از آن‌ها استخراج کرد و در آخر، با توجه به اینکه، داده‌های این پژوهش مربوط به یک کشت و صنعت است، پیشنهاد می‌شود پژوهشی مشابه، در سایر کشت و صنعت‌های نیشکری انجام گیرد و نتایج آن با نتایج این پژوهش و روش‌های دیگر، مقایسه شود.

سپاسگزاری

این مقاله مستخرج از طرح پژوهشی به شماره ۱۲۷۴ از محل اعتبارات پژوهانه واحد پژوهشی دانشگاه شهید چمران اهواز می‌باشد. بدین وسیله نویسندگان از معاونت پژوهشی دانشگاه شهید چمران اهواز بابت تأمین هزینه‌های این پژوهش سپاسگزاری می‌نمایند.

References

1. Baisen, Z., and R. Tillman. 2007. A decision tree approach to modeling nitrogen fertilizer use efficiency in New Zealand pasture. Plant and Soil 301 (1): 267-278.
2. Breman, L., Friedman, J., Olshen, R., and Ch. Stone. 1984. Classification and regression trees. Boca Raton: Chapman & Hall/CRC.
3. Choi, J., K. H. Jeon, Y. Won, and J. J. Kim. 2014. Pattern classification of foot diseases using decision tree. Wseas Transactions on Biology and Biomedicine 11: 157-164.
4. Ekasingh, B., and K. Ngamsomsuke. 2009. Searching for simplified farmers’ crop choice models for integrated watershed management in Thailand: A data mining approach. Environmental Modeling and Software 24: 1373-1380.
5. Geetha, M. C. S. 2015. A survey on data mining techniques in agriculture. International Journal of Innovative Research in Computer and Communication Engineering 3 (2): 887-892.
6. Goktepe, A. B., S. Altun, and A. Sezar. 2005. Soil clustering by fuzzy C-Means algorithm. Advances in Engineering Software 36: 691-698.
7. Jeysenthil, K. M. S., T. Manikandan, and V. Murali. 2014. Third generation agricultural support system development using data mining. International Journal of Innovative Research in Science, Engineering and Technology 3 (3): 9923- 9930.
8. Kalpana, R., N. Shanthi, and S. Arumugam. 2014a. Data mining– An evolutionary view of agriculture. International Journal of Application or Innovation in Engineering and Management 3 (3): 102- 105.
9. Kalpana, R., N. Shanthi, and S. Arumugam. 2014b. a survey on data mining techniques in agriculture. International Journal of Advances in Computer Science and Technology 3 (8): 426-431.
10. Kotsiantis, S. B. 2007. Supervised machine learning: A review of classification techniques. International Journal of Computing and Informatics 31 (3): 249- 268.
11. Monjezi, N., and H. Zakidizaji. 2017. Fuzzy approach to optimize overhaul time of sugarcane harvester using GERT network method. Iranian Journal of Biosystem Engineering 48 (1): 83-91. (In Farsi).
12. Monjezi, N., H. Zakidizaji, M. J. Sheikhdavoodi, A. Marzban, and M. Shomeili. 2017. Finding and prioritizing of effective parameters on lack of timeliness operations of sugarcane production using Analytical Hierarchy Process (AHP). Journal of Agricultural Machinery 7 (2): 514-526. (In Farsi).
13. Noorzadeh, M., K. Khavazi, M. Malakooti, and S. Hashemi. 2011. Evaluation of the effectiveness of C-means and GK methods for fuzzy clustering of copper concentration in agricultural lands (Case study: Hamedan Province). Journal of Agricultural Engineering 33 (1): 61-70. (In Farsi).
14. Quinlan, J. 1993. Programs for machine learning. Morgan Kaufmann, San Francisco, CA, pp.
15. Rajesh, D. 2011. Application of Spatial Data Mining for Agriculture. International Journal of Computer Applications 15 (2): 7-9.
16. Ramesh, D., and B. Vishnu Vardhan. 2013. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering 2 (9): 3477-3480.
17. Raorane, A. A., and R. V. Kulkarni. 2013. Review- Role of data mining in agriculture. International Journal of Computer Science and Information Technologies 4 (2): 270-272.
18. Razi Ardakani, H., and A. Samimi. 2011. Comparison of decision tree in modeling choosing a type of means of carriage of goods. 11th Transportation and Traffic Engineering. 2-3 February, Tehran. (In Farsi).
19. Sharma, L., and N. Mehta. 2012. Data mining techniques: A tool for knowledge management system in agriculture. International Journal of Scientific and Technology Research 1 (5): 67-73.
20. Umesh, D. R., and C. R. Thilak. 2015. Predicting breast cancer survivability using Naïve Baysien and C5.0 algorithm. International Journal of Computer Science and Information Technology Research 3 (2): 802-807.

21. Yethiraj, N. G. 2012. Applying data mining techniques in the field of agriculture and allied sciences. International Journal of Business Intelligents 1 (2): 72-76.
22. Yoneyama, Y., S. Suzuki, R. Sawa, K. Yoneyama, G. G. Power, and T. Araki. 2002. Increased plasma adenosine concentrations and the severity of preeclampsia. Obstet Gynecol 100 (6):1266-1270.

Journal of Agricultural Machinery
Vol. 9, No. 2, Fall - Winter 2019, p. 469-484

نشریه ماشین‌های کشاورزی
جلد ۹، شماره ۲، نیمسال دوم ۱۳۹۸، ص ۴۸۴-۴۶۹

Modeling of the Variables that Influence Sugarcane Yield using C5.0 and QUEST Decision Tree Algorithms

H. Zakidizaji[1*]- H. Bahrami[2]- N. Monjezi[3]- M. J. Sheikhdavoodi[4]

Received: 26-12-2017
Accepted: 07-04-2018

Introduction

The sugar industry usually gathers huge amounts of information during normal production operations, which is rarely used to study the relative importance of both management and environment on sugarcane yield performance. Yield prediction is a very significant problem of agricultural organizations. Each agronomist wants to know how much yield to expect as soon as possible. The aim of this study was to determine the performance of C5.0 and QUEST algorithms to predict the yield of sugarcane production in Amir-Kabir agro-industry Company of Khuzestan province, Iran. However, the working method described in this paper is applicable to other geographical areas and other kinds of crops.

Materials and Methods

The data for the study were collected from Amir-Kabir agro-industry Company. The data is obtained from 2012 to 2016 years. The study area is located in Khuzestan Province which is a major agricultural region in Iran. The geographical location of the study area is between latitudes 31° 15′ to 31° 40′ north and longitudes 48° 12′ to 48° 30′ east. It covers an area of about 12000 ha. The average elevation of the study area is 8m above sea level. Mean annual rainfall within the study area is 147.1mm, the mean annual temperature is approximately 25°C and the mean soil temperature at 50cm depth is 21.2°C. The used data were obtained from a survey with 15 variables carried out on 1201 sugarcane farms. Variables used in the study of data mining can be divided into two categories: target variable and predictor variables. The variable of yield was used as the target variable (dependent) and other variables as predictor variables (independent). In two models, the input data included crop cultivar, month of harvest, chemical fertilizer (Nitrogen), chemical fertilizer (Phosphate), age (plant or ratoon), times irrigation, ratio of surface spraying, soil texture, soil electrical conductivity (EC), water consumption per hectare, drain, farm management, crop duration, area, and yield-category. The study was included in 1201 farms. The necessary data were collected and pre-processing was performed. We propose to analyze different decision tree methods (C5.0 and QUEST).

Results and Discussion

First, decision tree methods were analyzed for variables. Then, according to C5.0 method (error rate 0.2319 for the training set and 0.3306 for test set) performed slightly better than another method in predicting yield. Crop cultivar is found that an important variable for the yield prediction. 24 rules were found in this study, C4.5 showed a better degree of separation. The measured prediction rate of C5.0 was correct: 76.81% and wrong: 23.19% in the training data, and correct: 66.94% and wrong: 33.06% in the test data. The prediction rate of QUEST was correct: 68.25% and wrong: 31.75% in the training data, and correct: 70.83% and wrong: 29.17% in the test data. Using the training data comparison between the model types showed that the C5.0 model produces a more accurate prediction model and was, therefore, the model to use. Using the testing data in comparison with the model types showed that the QUEST model produced a more accurate prediction model. The results of our assessment showed that C5.0 and QUEST algorithms were capable to produce rules for sugarcane yield. Therefore, our proposed methods as an expert and intelligent system had an impressive impact on sugarcane yield prediction.

Conclusions

In today's conditions, agricultural enterprises are capable of generating and collect large amounts of data.

1- Assistant Professor, Biosystems Engineering Dept., Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran
2- Associate Professor, Biosystems Engineering Dept., Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran
3- Assistant Professor, Biosystems Engineering Dept., Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran
4- Professor, Biosystems engineering Dept., Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran
(*- Corresponding Author Email: hzakid@scu.ac.ir)

Growth of data size requires an automated method to extract necessary data. By applying data mining technique it is possible to extract useful knowledge and trends. Knowledge gained in this manner may be applied to increase work efficiency and improve decision making quality. Data mining techniques are directed towards finding those schemes of work in data which are valuable and interesting for crop management. In this research, decision tree algorithms (C5.0 and QUEST) were used. This classification algorithm was selected because it has the potential to yield good results in prediction and classification applications. This study was performed to present a model-based data mining to predict sugarcane yield in 2012-2016. The 24 classification rules generated from the C5.0 decision tree algorithm have great practical value in agricultural applications. The results showed the QUEST and C5.0 decision tree algorithms produced the best prediction accuracy. Sensitivity analysis results indicated that crop cultivar was the most important variables. It was observed that efficient technique can be developed and analyzed using the appropriate data, which was collected from Khuzestan province to solve complex agricultural problems using data mining techniques (decision tree). The decision tree has been found useful in classification and prediction modeling due to the fact that it can capability to accurately discover hidden relationships between variables, it is capable of removing insignificant attributes within a dataset.

Keywords: Agriculture, Amir-Kabir Agro-Industry, Data mining, Evaluation